学校编码：10384　　分类号_____密级_____
学号：24320071151834　　UDC_____

厦门大学

硕 士 学 位 论 文

电子商务客户流失分析技术与模型研究

Research on Technologies and Models for Customer Churn Analysis in Electronic Commerce

刘菡

指导教师姓名：董 槐 林　　教授
专 业 名 称：计算机软件与理论
论文提交日期：2010　年　5　月
论文答辩时间：2010　年　5　月
学位授予日期：2010　年　　月

答辩委员会主席：_____
评　　阅　　人：_____

2010 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘　要

电子商务的普及带来了电子商务企业之间的激烈竞争，客户资源是最大的竞争目标。随着各类购物网站的兴起，客户流失不可避免。在传统客户关系管理的基础上，结合电子商务高度交互的特点，分析和预测客户流失已成为电子商务客户管理的一个趋势。由于电子商务交易涉及多个操作过程，因此产生了大量数据，如客户基本资料、点击流数据、交易记录等。数据挖掘能够从大量数据中筛选出有用的、隐含的信息，并揭示数据之间存在的潜在联系，比如客户的网购经验与取消订单次数之间的关系，因此适合应用于客户的行为分析。目前，应大规模数据的处理需求，已产生了多种数据挖掘技术，但将这些基本算法应用于特殊领域时还需要在算法效率上进行改进。

论文工作针对目前电子商务客户流失日益严重的问题，在分析已有客户流失分析方法的基础上，对电子商务客户流失分析的技术及模型进行了系统、深入和较为全面的研究。这些研究内容不仅是电子商务客户管理亟待解决的关键问题，同时也是数据挖掘领域的研究重点，具有重要的理论意义和实际应用价值。论文的主要工作和贡献如下：

(1) 结合抽样理论，对 DBSCAN 算法进行改进。引入均匀抽样，并利用随机抽样标记数据点以减少数据对象的查询次数，提出二次抽样 DBSCAN 方法（DSDBSCAN）。实验表明该算法适用于大规模数据集的聚类，并在不影响聚类效果的前提下，加快数据处理的速度。

(2) 根据卡方统计量离散化数值属性的特点，结合 CART 算法，提出 χ^2-CART 算法，弥补了 CART 算法对数值属性二元划分时消耗大量运行时间的缺陷，并用于建立客户流失分析模型。

(3) 以一个网店的客户数据为研究对象，利用 DSDBSCAN 算法，根据客户的当前价值和潜在价值对客户数据进行聚类，然后对高潜在价值客户的行为模式进行分析，通过 χ^2-CART 算法生成的决策树，找出客户流失的规律，预测哪些客户具有流失倾向，并归纳模型。

关键词：客户流失；DSDBSCAN；χ^2-CART

Abstract

With the wide application of electronic commerce (EC), drastic competition has been brought to EC enterprises. Customers are the biggest competitive resource and customer churn is inevitable with the development of kinds of on-line malls. As a result, analysis and prediction of customer churn has been a tendency in customer relationship management of EC. EC involves several phases which generate large-scale data, including personal information of customers, click-streams and transaction records. Data mining can screen out useful and connotative information, as well as potential relations, for example, the relation between customers' on-line shopping experience and the times of canceling orders. Therefore, data mining is suitable for analyzing consumers' behaviors. At present, a certain data mining technologies have come into being to satisfy the requirements of handling large-scale data, but basic algorithms need to be improved when applied to specific areas.

As customer churn in EC has become a serious problem, this dissertation makes systematical and deep research on technologies and models for customer churn in EC, based on existed analysis methods of customer churn. These researches are not only key problems to be solved in EC, but also emphases in the field of Data Mining. The main work is as follows:

(1) Sampling theory is applied to DBSCAN algorithm to generate a new approach, Double-Sampling DBSCAN. Experiments show that it can expedite the speed of clustering on large-scale data without affecting the effect of aggregation. During the process, well-proportioned sampling is introduced and randomly labeling data points can reduce the query times of data objects.

(2) Combined HL Chi Square and CART algorithm, χ^2-CART is proposed to construct the model for analyzing and predicting customer churn. The new approach solves low time efficiency problem existed in CART, which is caused by exponential computations generated by dual partitions on numerical attributes, as HL Chi Square can disperse numerical attributes.

(3) Taking customer data extracted from a network shop as the research object, DSDBSCAN is applied to clustering according to current value and potential value of customers; Then through analyzing the behavior mode of customers with high potential value, and generating decision trees using χ^2-CART, rules of customer churn is summarized. Finally, the model for analyzing and predicting customer churn is depicted to predict which kinds of customer have the inclination to churn.

Key Words: Customer Churn; DSDBSCAN; χ^2-CART

目录

第一章 绪论 1

1.1 研究背景及意义 1

1.2 国内外研究现状 2

1.3 论文的主要工作 4

1.4 论文内容结构 6

第二章 电子商务客户流失理论及相关技术分析 7

2.1 电子商务客户的研究价值 7

2.1.1 电子商务客户的发展背景 7

2.1.2 客户价值 8

2.1.3 聚类方法细分客户 10

2.2 客户流失相关理论 11

2.2.1 客户忠诚度 12

2.2.2 客户流失 14

2.3 客户流失的挖掘方法 16

2.3.1 数据挖掘概述 16

2.3.2 数据挖掘的主要算法 18

2.3.3 客户聚类和流失分析建模的算法选择 19

2.4 本章小结 21

第三章 二次抽样 DBSCAN 算法研究 22

3.1 聚类分析概述 22

3.1.1 聚类的概念 22

3.1.2 聚类方法分类 23

3.1.3 客户聚类算法的要求 26

3.2 DBSCAN 算法 28

3.2.1 DBSCAN 算法思想 28

3.2.2 DBSCAN 算法局限性 29

3.3 二次抽样 DBSCAN 30

3.3.1 抽样理论概述.....30
3.3.2 DSDBSCAN 算法介绍.....32
3.3.3 DSDBSCAN 算法性能分析.....35
3.4 本章小结.....38
第四章 χ^2-CART 算法研究.....39
4.1 决策树概述.....39
4.2 决策树建模算法的研究.....41
4.2.1 ID3 算法.....41
4.2.2 C4.5 算法.....43
4.2.3 CART 算法.....45
4.2.4 决策树建模算法比较.....46
4.3 χ2-CART 算法.....48
4.3.1 卡方统计量概念.....48
4.3.2 χ^2-CART 算法介绍.....49
4.3.3 χ^2-CART 算法性能分析.....51
4.4 本章小结.....52
第五章 电子商务客户流失分析模型建立.....53
5.1 数据预处理.....53
5.1.1 数据采集.....53
5.1.2 数据清理.....55
5.1.3 数据变换.....56
5.1.4 属性子集选择.....57
5.2 建立流失分析模型.....58
5.2.1 客户特征聚类.....58
5.2.2 DSDBSCAN 实施客户聚类.....59
5.2.3 χ^2-CART 分析客户流失.....61
5.2.4 归纳客户流失分析模型.....63
5.3 客户挽留措施.....65
5.3.1 流失原因分析.....65

5.3.2 挽留决策制定.....66
5.4 本章小结.....67
第六章 总结与展望.....68
6.1 总结.....68
6.2 展望.....69
参考文献.....70
攻读硕士期间科研成果.....76
致谢.....77

Contents

Chapter 1 Introduction.....1

1.1 Background and Significance 1

1.2 Current Research Status 2

1.3 Main Work.....4

1.4 Framework 5

Chapter 2 Analysis of Theories and Technologies of Customer Churn in Electronic Commerce 7

2.1 Research Value of EC Customers.....7

2.1.1 Development Background 7

2.1.2 Cusomer Value 8

2.1.3 Segmentation of Customers Using Clustering.....10

2.2 Theories of Customer Churn 11

2.2.1 Customer Loyalty.....12

2.2.2 Customer Churn 14

2.3 Mining Methods of Customer Churn.....16

2.3.1 Introduction of Data Mining 16

2.3.2 Main Data Mining Technologies.....18

2.3.3 Algorithms Choosing for Customer Clustering and Modeling 19

2.4 Summary.....21

Chapter 3 Research on Double-Sampling DBSCAN.....22

3.1 Clustering Algorithm 22

3.1.1 Conception of Clustering 22

3.1.2 Classification of Clustering.....23

3.1.3 Requirements of Customer Clustering.....26

3.2 DBSCAN Algorithm 28

3.2.1 Definition of DBSCAN.....28

3.2.2 Deficiency of DBSCAN 29

3.3 Double-Sampling DBSCAN 30
3.3.1 Sampling Theory 30
3.3.2 Introduction of DSDBSCAN 32
3.3.3 Efficiency Analysis of DSDBSCAN 35
3.4 Summary 38
Chapter 4 Research on χ^2-CART 39
4.1 Definition of Decision Tree 39
4.2 Classification of Decision Tree Algorithm 41
4.2.1 ID3 41
4.2.2 C4.5 43
4.2.3 CART 45
4.2.4 Comparation of Decision Tree Algorithms 46
4.3 χ^2-CART Algorithm 48
4.3.1 HL Chi Square 48
4.3.2 Introduction of χ^2-CART 49
4.3.3 Efficiency Analysis of χ^2-CART 51
4.4 Summary 52
Chapter 5 Model Construction for Customer Churn Analysis in EC 53
5.1 Data Preprocessing 53
5.1.1 Data Collection 53
5.1.2 Data Cleaning 55
5.1.3 Data Transition 56
5.1.4 Attribute Subsets Choosing 57
5.2 Model Construction 58
5.2.1 Customer Feature Clustering 58
5.2.2 DSDBSCAN Applied to Clustering 59
5.2.3 χ^2-CART Applied to Analyze Customer Churn 61
5.2.4 Analysis and Prediction Model 63

5.3 Strategies for Customer Detainment **65**
5.3.1 Reasons for Customer Churn 65
5.3.2 Detainment Decision 66
5.4 Summary **67**
Chapter 6 Conclusions and Prospects **68**
6.1 Conclusions **68**
6.2 Future Work **69**
References **70**
Publications **76**
Acknowledgements **77**

第一章 绪论

激烈的市场竞争带动电子商务水平的不断提高，客户已成为各电子商务企业最重要的竞争对象。然而，在有限的市场中面对众多的不断变化的竞争对手和日益更新的技术，客户流失是不可避免的。如果不对客户数据加以分析，及时采取应对措施，这种流失将导致电子商务企业的利润下降。

1.1 研究背景及意义

近十年来，因特网的发展促进了电子商务的迅速崛起，网络购物已成为广大消费者购物的一个重要途径。中国互联网络信息中心(CNNIC)对国内的网络购物行为的调查结果[1]显示，2006 年和 2007 年是中国电子商务迅速崛起并高速发展的时期；截止到 2009 年 6 月份，全国网民突破 3 亿，其中网络购物的人数占约三分之一；与此同时，网络购物的交易额也快速上升，2009 年上半年，国内网络购物的市场规模突破千亿，相比 2008 年上半年，同比增长了 94.8%[2]。

电子商务的发展成为趋势，意味着更多的企业或个人加入了竞争的行列。大部分商家注重采用创新或推荐策略吸引新客户，而忽视了老客户的重要性。研究发现[3]，吸引一个新客户所用的成本远远高于维系一个老客户的成本(Reicheld and Sasser, 1990)。由此看来，分析客户流失的原因，预测可能流失的客户，并制定相应的决策防止客户流失具有实际研究意义。

由于电子商务网络交易功能的完整性，除了商品的运送是在线下完成之外，商家与客户双方之间的接触、交易的确定及资金流动均在网上完成。由此产生的点击流数据、交易数据及其他相关信息的规模远远超过了传统产业，此时使用传统的客户分组方法[4]来分析和处理这类大规模数据集是不可行的。因此，如何有效地处理大规模客户信息，挖掘出导致客户流失的因素，预测潜在流失的客户成为电子商务商家面临的一大问题。

数据挖掘技术的出现和发展为该问题的解决提供了强有力的工具。数据挖掘技术伴随着处理大规模数据的市场需求的增长而不断发展，可以说，数据挖掘技术应用于电子商务客户的分析几乎是从数据挖掘一出现就注定的。然而，面对如此具体的应用背景和领域，传统的数据挖掘算法在处理大规模电子商务客户数据

方面仍存在诸多局限。因此在对客户分组时需要对传统的数据挖掘算法（如聚类算法）进行改进；此外，还考虑将不同数据挖掘算法有效结合，如融合密度聚类算法[5]和决策树[6]算法，以提高客户流失的预测能力。

基于上述分析，本文将围绕电子商务环境下客户流失分析这一主题，通过分析客户的基本数据、交易数据和行为模式，着重从时间效率角度对聚类算法进行改进以快速挖掘出有用信息，并利用决策树建立客户流失分析模型，在此基础上帮助企业分析完成任务所需要的关键因素，以降低成本并增加收入，使电子商务商家处于更有利的竞争位置。

1.2 国内外研究现状

国外对这一方面的研究已有八、九年的研究时间，其应用领域主要是电信业、金融业和保险业等行业，采用的方法包括：

1. 统计分析方法

统计分析方法[7]主要用于关系型知识挖掘[8]，找出统计分析关系表中各属性之间存在的关系。一般来说，属性之间的关系有两种：一种是函数关系，即能用函数公式表示的确定性关系。另一种是相关关系，也就是不能用函数公式表示的关系。比如在人的年龄与网购经验，两个变量之间虽然存在着密切的关系，但却不能由一个或几个变量的数值精确地求出另一个变量的值。但由于测量误差，确定性关系往往可以通过相关关系呈现出来；在了解事物内部规律的情况下，相关关系也有可能转化成确定性关系，对它们可采用相关分析及主成分分析等方法。

2. 遗传算法

遗传算法[9]主要用于分类和关系型规则挖掘。遗传算法从一个初始的规则集合开始，通过交换对象产生群体，采用优胜劣汰的规则进行评估，逐步积累计算，以得到最优化的知识集。在算法的执行过程中，存在着同一代的许多不同的种群个体，这些个体的保留和淘汰，是由它们对环境的适应能力决定的。适应能力通过计算适应度函数 f(x)的值进行判别,这个值称为适应值。实际上，适应度函数 f(x)通常是目标函数的变种。经过计算，适应能力强的个体具有更大的机会保留下来。遗传算法具有自适应、自组织和自学习性的特点，采用概率的变化规则指导其搜索方向。

3．决策树方法

决策树方法主要用于分类。决策树算法利用信息论中的信息增益搜索具有最大信息增益的属性，建立根节点，然后以某种属性度量选择方法为依据将原始样本集划分为若干个分支子集。对每个分支子集重复上述步骤，即可生成一棵决策树。决策树的根节点到叶节点的每一条路径都能够通过各属性分支的合取生成一条分类规则，整棵决策树就是所有分类规则的集合。生成一棵决策树之后，可以对其进行剪枝处理并将其转化为规则，然后对新样本进行分类。决策树算法主要包括 CLS、ID3、C4.5 及 CART 等。

4．人工神经网络

人工神经网络[10]主要用于群集、分类、预测、模式识别和特征挖掘。人工神经网络的原理模仿了生物神经网络，其本质上是一个分散型或矩阵结构，通过挖掘训练数据，逐步计算网络连接的加权值。神经网络中处理单元的类型可分为三类，分别是输入单元、隐单元和输出单元。输入单元接受外部信号与数据，输出单元用于输出系统的处理结果，隐单元是不能由系统外部观察的单元，处于输入和输出单元之间。神经单元间的连接强度是由神经元之间的连接权值所反映的，信息的表示和处理也体现在单元之间的连接关系中。实际上，人工神经网络是通过动力学行为和网络的变换得到一种并行分布式的、在不同程度和层次上模仿人脑神经系统的信息处理功能，是一种非程序化、适应性及大脑风格的信息处理。

5．粗糙集

粗糙集[11]主要用于数据简化、因果关系、对象相似性分析、数据意义评估及范式挖掘等。粗糙集理论由 Z.Pwa lka 在 20 世纪 80 年代提出，是知识发现的一种重要的研究方法，它所采用的信息表类似于关系数据库中的关系数据模型，主要用于处理不确定性。粗糙集方法能处理各种数据，包括不完整数据或具有众多变量的数据；它能处理数据的不精确性。此外，粗糙集还可以求得知识的最小表达和各种不同颗粒层次；能从数据中揭示出概念简单,易于操作的模式；还能产生精确的检查和证实的规则，适用于智能控制中规则的自动生成。

6．联机分析处理技术

联机分析处理[12](On Line Analysis Processing，OLAP)主要用于针对特定问题的联机数据访问和分析。OLAP 技术以多维数据库为基础，对数据转化成信

息或知识非常有用，通过对信息多种可能的观察形式能够进行快速、稳定一致的和交互性的存取，并允许管理决策人员深入观察数据。OLAP 专门设计用于支持复杂的分析操作，可根据分析人员的要求进行快速灵活地大数据量的复杂查询处理，并提供一种直观易懂的查询结果给决策人员，从而帮助他们准确了解对象的需求，掌握企业的经营状况，制定出正确的方案。OLAP 还可用于证实人们提出的复杂的假设，以图形或者表格表示对信息的总结。此外，OLAP 技术在挖掘系统中融入了人的观察力和智力，提高了系统挖掘的速度。

国外某些科研机构已经研究并提出了较为成熟的客户流失分析模型[13,14]，但从市场的反馈来看，这些模型并不具备很强的健壮性,准确率也不是很高。而且，随着数据量的激增，对模型的性能开销也越来越大，所以需要提出改进的算法对客户流失分析模型进行优化和完善。

目前国内对电子商务客户关系的研究侧重于提供个性化服务，如个性化推荐以吸引客户，提高客户忠诚度；很少直接对客户流失数据进行建模，分析客户流失原因，对潜在流失客户预测，以采取有效防范措施。

1.3 论文的主要工作

客户流失分析首先将客户分组，然后对具有研究价值的客户群体加以分析。传统方法一般是基于经验的分类方法，由决策者根据以往的经验划分客户，因此具有较强的主观性；另一种常用方法是统计方法，简单统计客户的属性特征并以此划分客户。虽然这些划分对客户管理也是具有一定意义的，但不能满足一些复杂分析的需求，比如影响客户流失的因素、客户流失的概率、不同类别客户的流失情况的差异情况。此外，数据库技术的应用使得电子商务企业积累了大量的数据，所以企业往往希望从大量的数据中提取有用的知识，建立起有效的客户流失分析模型，因此数据挖掘中的聚类和分类预测技术应需求而产生。

数据挖掘[15]是指从数据库或其他存储介质中提取出人们感兴趣的，事先未知的，有用的或潜在有用的信息。主要方法包括聚类分析，关联分析，分类和预测，孤立点分析和演变分析。作为数据挖掘的主要任务之一，聚类分析能够获得数据的分布状况，观察每一个簇中数据的特征，并集中对特定的簇进一步分析。除了广泛应用于商业、生物学等领域，聚类分析还可以作为其他算法的预处理步骤。